GRUPO 10 - PRÁCTICA PROFESIONALIZANTE 1 - ISPC

CREACIÓN DE UN MODELO DE MACHINE LEARNING UTILIZANDO UN DATASET CON LAS LLAMADAS DE DENUNCIA DE CASOS VIOLENCIA AL 144 EN PROVINCIA DE BSAS - CABA Y RESTO DEL PAÍS.

INTEGRANTES:

Quiroga Horacio

Bustos Jonathan

Metz Claudia

Hilgemberg María Sol

Meier Iván Didier

Muñoz Mariel

Soria Julio Ezequiel

Flores Nadia Daniela

-2023-

BASE DE DATOS DE LA LÍNEA 144 MINISTERIO DE LAS MUJERES, GÉNEROS Y DIVERSIDAD DE LA NACIÓN - DIRECCIÓN TÉCNICA DE REGISTROS Y BASES DE DATOS.

RECURSOS:

CSV >>> https://datos.gob.ar/dataset/generos-base-datos-linea-144

COLAB >>> *https://colab.research.google.com/drive/1zxKEKvCfkHKwMbdQOWSFJMwOQVNWZRX3#scrollTo=llzJf-FL8WQL*

TRELLO >>> https://trello.com/b/SE0rBeCR/tablero-para-el-trabajo-pr%C3%A1ctico-del-grupo-10-de-pr%C3%A1ctica-profesionalizante-1-2023

GITHUB >>> https://github.com/ClaudiaMetz/ISPC---PP1---Grupo-10

EL PROYECTO

Decidimos utilizar un dataset que contiene los datos de las denuncias de diversos tipos de violencia (en su inmensa mayoría de mujeres hacia hombres) formuladas en llamados a la línea de Denuncias 144 que es de cobertura nacional. La información es desde el 2020 al 2023 (1er Semestre)

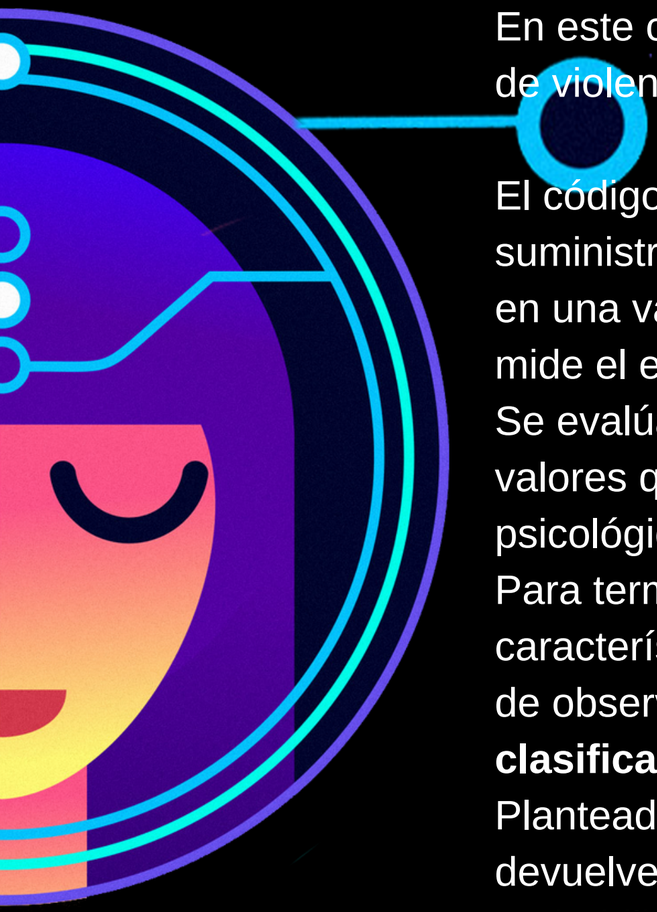
ETAPA EXPLORATORIA

Analizamos los datos del dataset que elegimos. Hicimos varias gráficas para observarlos mejor.
Renombramos columnas, completamos los pocos datos faltantes, cambiamos el formato de datos a otros que sean coherentes con los necesarios para idear el modelo de Machine Learning.

EL MODELO

Dado que los datos con los que contamos son en su mayoría binarios (si - no) nos surgió la idea de plantear un modelo de Regresión Logística y también uno de Random Forest.

REGRESIÓN LOGÍSTICA
CASO 1



En este caso, el modelo intenta predecir si es probable que una persona haya denunciado que fue víctima de violencia psicológica o no, en función de su edad.

El código usa el método *fit* para entrenar el modelo con los datos de entrenamiento que le hemos suministrado y el método *predict* para hacer predicciones sobre los datos de prueba guardando el resultado en una variable. El método *fit* ajusta los parámetros del modelo para minimizar una función de pérdida que mide el error entre las predicciones y las etiquetas reales.
Se evalúa el modelo con los datos de prueba usando el método *predict*. Las predicciones contienen 2 valores que indican si el modelo clasifica a la persona propuesta como posible víctima o no de violencia psicológica.
Para terminar, se calcula la precisión del modelo usando el método *score*, que recibe como parámetros las características y las etiquetas de prueba y devuelve la proporción de predicciones correctas sobre el total de observaciones. **La precisión es una medida de rendimiento del modelo que indica qué tan bien clasifica a las víctimas.**
Planteado todo esto, podemos hacer predicciones para casos particulares. En este caso, el método devuelve un array con dos valores: la probabilidad de que la víctima no sufra violencia psicológica (0) y la probabilidad de que sí la sufra (1).

REGRESIÓN LOGÍSTICA
CASO 1

Interpretación de Resultados

La salida [0.06659837 0.93340163] indica que el modelo ha predicho que hay una probabilidad del 6.66% de que la persona que hizo el llamado no haya sufrido violencia psicológica (ViolenciaPsicologica = 0) y una probabilidad del 93.34% de que sí haya sufrido violencia psicológica (ViolenciaPsicologica = 1). A su vez, la precisión del modelo es bastante buena: 94,63%

REGRESIÓN LOGÍSTICA
CASO 2

Es muy similar al anterior, solo que tiene una diferencia: En este caso, se usan dos características para entrenar el modelo: la edad de la víctima y si sufrió o no violencia física. Esto significa que el modelo trata de estimar la probabilidad de que una víctima sufra violencia psicológica en función de su edad y de si fue agredida físicamente o no. Para ello, se selecciona un subconjunto del DataFrame con las columnas ‘EdadVictima’ y ‘ViolenciaFisica’. La etiqueta sigue siendo la misma: la variable ‘ViolenciaPsicologica’. El resto del código es igual al caso anterior, se divide el conjunto de datos en entrenamiento y prueba, se crea el modelo de regresión logística, se entrena el modelo con los datos de entrenamiento, se evalúa el modelo con los datos de prueba, se calcula la precisión del modelo y se hace una predicción para un caso particular.

REGRESIÓN LOGÍSTICA
CASO 2

Interpretación de Resultados

La salida [0.04071707 0.95928293] indica que el modelo ha predicho que hay una probabilidad del 4.05% de que la víctima no haya sufrido violencia psicológica (ViolenciaPsicologica = 0) y una probabilidad del 95.95% de que sí haya sufrido violencia psicológica (ViolenciaPsicologica = 1) habiendo sufrido a la vez violencia física.

REGRESIÓN LOGÍSTICA
CASO 3

En este caso se usan tres características para entrenar el modelo: la edad de la víctima, si sufrió o no violencia psicológica y su residencia. El modelo trata de estimar la probabilidad de que una víctima sufra violencia doméstica en función de su edad, de si fue agredida psicológicamente o no y teniendo en cuenta la provincia donde radica.

Para ello, se selecciona un subconjunto del DF con las columnas ‘EdadVictima’, ‘ViolenciaPsicologica’ y 'label_encoder' que contiene el valor numérico asignado a cada provincia.

La etiqueta es la variable ‘ViolenciaDomestica’. Se divide el conjunto de datos en entrenamiento y prueba, se crea el modelo de RL, se entrena el modelo con los datos de entrenamiento, se evalúa el modelo con los datos de prueba, se calcula la precisión del modelo y se hace una predicción para un caso particular.

REGRESIÓN LOGÍSTICA
CASO 3

Interpretación de Resultados

La salida [0.07058691 0.92941309] indica que el modelo ha predicho que hay una probabilidad de 7.06% que una persona de 40 años, que sufrió violencia Psicológica y es de Córdoba, no haya sufrido violencia domestica (ViolenciaDomestica = 0) y una probabilidad del 92.94% de que sí haya sufrido violencia Domestica (ViolenciaDomestica = 1).

La violencia psicológica y la violencia física son dos formas diferentes de violencia, pero a menudo están relacionadas entre sí. La violencia psicológica puede ser un precursor de la violencia física, y la violencia física puede tener efectos psicológicos duraderos en las víctimas.

En Argentina, la Ley 26.485 define la violencia contra las mujeres como "cualquier conducta que ataque: Tu vida. Tu libertad. Tu dignidad. Tu integridad física, psicológica o sexual. Tu situación económica. Tu seguridad personal. Tu participación política". *La ley reconoce que la violencia contra las mujeres puede tomar muchas formas diferentes, incluyendo la violencia física, la violencia psicológica, la violencia sexual, la violencia económica y la violencia simbólica.*

Un estudio realizado en Argentina encontró que las mujeres que habían experimentado al menos una forma de violencia psicológica tenían casi 10 veces más probabilidades de ser víctimas de violencia física o sexual por parte de su pareja actual. Esto sugiere que la violencia psicológica y la violencia física a menudo ocurren juntas.

RANDOM FOREST

Usamos el método *dt* para extraer el año, el mes y el día de la fecha y los guardamos en nuevas columnas llamadas año, mes y día. Luego, el código usa la función *pd.get_dummies* para realizar una codificación ***one-hot*** de la columna ResidenciaVictima, que indica el lugar donde vive la víctima. La codificación *one-hot* consiste en crear una columna binaria por cada valor único de la columna original, y asignar un 1 si la fila tiene ese valor, o un 0 si no lo tiene. Esto se hace para facilitar el uso de variables categóricas en modelos de aprendizaje automático. El código también incluye las columnas año y mes en el resultado de la codificación *one-hot,* y lo guarda en una variable llamada "X". La variable "y" se asigna a la columna ViolenciaFisica, que indica si la víctima sufrió violencia física o no.

A continuación, el código crea una instancia del *modelo RandomForestClassifier*, que es un algoritmo de aprendizaje automático supervisado que usa varios árboles de decisión para clasificar los datos. El parámetro *n_estimators* indica que se usarán 100 árboles, y el parámetro *criterion* indica que se usará la *entropía* como medida de impureza para dividir los nodos de los árboles. Finalmente, el código usa varias funciones para medir el desempeño del modelo: *accuracy_score* calcula la precisión del modelo, que es la proporción de predicciones correctas sobre el total; *confusion_matrix* muestra una matriz que compara las predicciones con los valores reales; y *classification_report* muestra un reporte con varias métricas como la precisión, la exhaustividad y la puntuación F1 por cada clase.

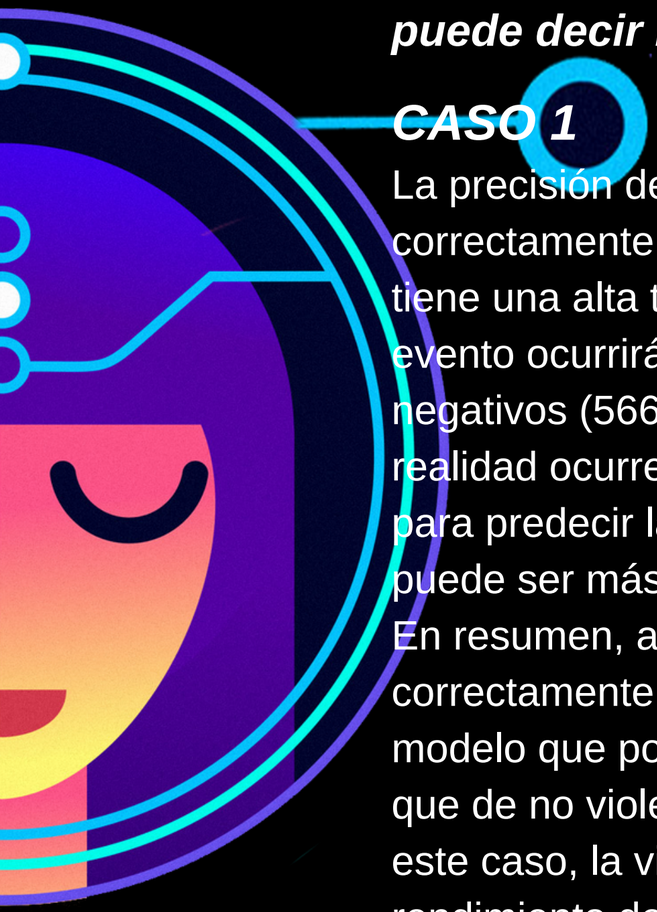
RANDOM FOREST -CASO 1

En el primer caso se trató de realizar un modelo de aprendizaje automático que utiliza el algoritmo de Random Forest para predecir si una persona que hace la denuncia al 144 ha sufrido o no violencia física, en función del año y mes de la denuncia y su lugar de residencia.

RANDOM FOREST -CASO 2

En el segundo caso se trató de realizar un modelo de aprendizaje automático que utiliza el algoritmo de Random Forest para predecir la residencia de la víctima en función del año, el mes y si sufrió o no violencia física.

RANDOM FOREST

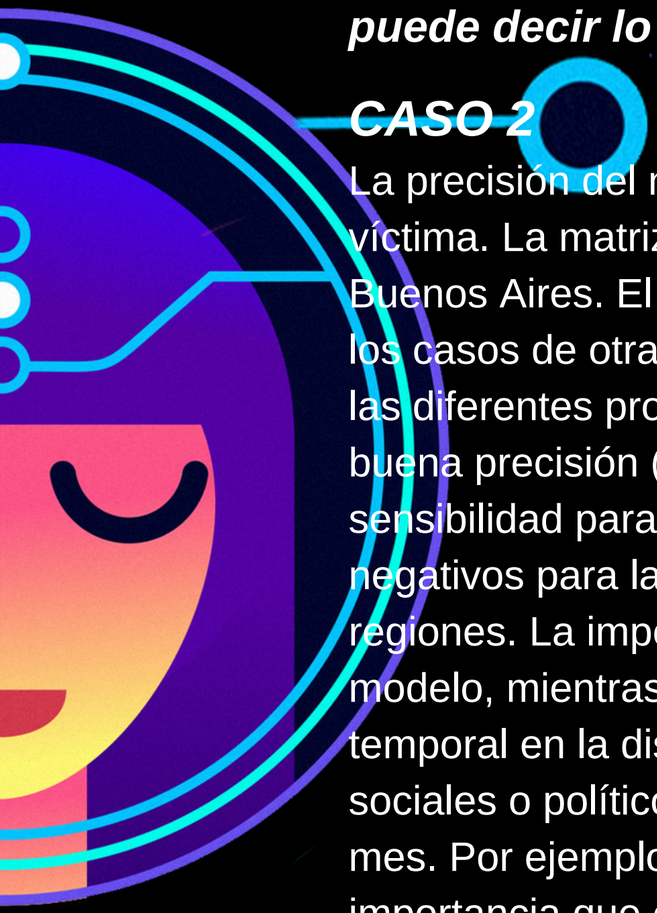


La interpretación del resultado depende del objetivo y el contexto del análisis, pero en general se puede decir lo siguiente:

CASO 1

La precisión del modelo es de aproximadamente 0.6516, lo que indica que el modelo es capaz de identificar correctamente el resultado positivo en un 65,16% de los casos y la matriz de confusión muestra que el modelo tiene una alta tasa de falsos positivos (5676 casos), lo que indica que el modelo predice incorrectamente que el evento ocurrirá en muchos casos donde no ocurre. Sin embargo, el modelo tiene una baja tasa de falsos negativos (566 casos), lo que indica que raramente predice incorrectamente que el evento no ocurrirá cuando en realidad ocurre. El reporte de clasificación muestra que el modelo tiene una precisión relativamente alta (0.66) para predecir la clase 1, pero una precisión más baja (0.48) para predecir la clase 0. Esto sugiere que el modelo puede ser más efectivo para predecir la ocurrencia del evento que para predecir su no ocurrencia.
En resumen, aunque el modelo tiene una precisión general decente, parece tener dificultades para predecir correctamente cuando el evento no ocurrirá. Esto podría ser un área a considerar para mejorar el rendimiento del modelo que podría deberse a un desequilibrio en los datos, donde hay muchos más ejemplos de violencia física que de no violencia física. En tales casos, los modelos tienden a estar sesgados hacia la clase mayoritaria (en este caso, la violencia física). Es posible que se necesiten técnicas de balanceo de clases para mejorar el rendimiento del modelo en la predicción de no violencia física.

RANDOM FOREST



La interpretación del resultado depende del objetivo y el contexto del análisis, pero en general se puede decir lo siguiente:

CASO 2

La precisión del modelo es de 0.55, lo que significa que acierta el 55% de las veces al predecir la residencia de la víctima. La matriz de confusión muestra que el modelo tiene un alto sesgo hacia la clase más frecuente, que es Buenos Aires. El modelo predice correctamente 9898 casos de Buenos Aires, pero clasifica incorrectamente todos los casos de otras provincias asimilándolos como Buenos Aires. Esto implica que el modelo no puede distinguir entre las diferentes provincias y siempre elige la más probable. El reporte de clasificación muestra que el modelo tiene una buena precisión (precision) y sensibilidad (recall) para la clase de Buenos Aires, pero una baja precisión y sensibilidad para las demás clases. Esto implica que el modelo tiene un alto costo de falsos positivos y falsos negativos para las otras provincias, lo que puede ser problemático si se quiere identificar a las víctimas de otras regiones. La importancia de las características muestra que el año y el mes son los factores más relevantes para el modelo, mientras que la violencia física tiene una menor influencia. Esto puede deberse a que hay una variación temporal en la distribución de las denuncias por provincias, que puede estar relacionada con factores estacionales, sociales o políticos. El gráfico de barras muestra que la importancia de las características varía según el año y el mes. Por ejemplo, en 2020, el mes tiene más importancia que el año, mientras que en 2023, el año tiene más importancia que el mes. El modelo predice que para el mes de julio de 2023, si una persona sufrió violencia física, hay un 64% de probabilidad de que su residencia sea Buenos Aires y un 36% de probabilidad de que sea CABA.

CONCLUSIÓN FINAL

En los modelos que hemos planteado ha sido mucho más eficiente utilizar la REGRESIÓN LOGÍSTICA que RANDOM FOREST